

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2003-316589

(43)Date of publication of application : 07.11.2003

(51)Int.Cl.

G06F 9/46
G06F 3/06

(21)Application number : 2002-120216

(71)Applicant : HITACHI LTD

(22)Date of filing : 23.04.2002

(72)Inventor : NAGASUGA HIROFUMI

KIYOI MASAHIRO

OTSUJI AKIRA

IKEGAYA NAOKO

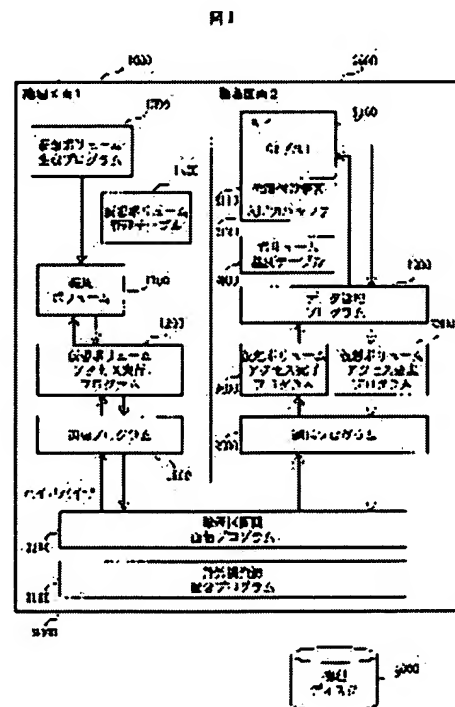
HIRAIWA YURI

(54) REAL MEMORY USING METHOD

(57)Abstract:

PROBLEM TO BE SOLVED: To solve the problem that delay due to the concentration of access is generated, or that the change of a user program is necessary for adopting a method for preventing the generation of the delay in a system where a shared data is disposed in an external storage device.

SOLUTION: A selected arbitrary computer in a virtual computer system is provided with a step for converting an address designated in an input/output request command transmitted from another computer to an external storage device by a communicating means into an address in a corresponding data region in a real storage device held in the selected logical block, a step for, when the command is an input request, returning the data stored in the data region of the converted address to a computer being the origin of transmission of the command, and a step for, when the command is an output request, storing the data transmitted with the command in the data region of the converted address.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2003-316589

(P2003-316589A)

(43) 公開日 平成15年11月7日 (2003.11.7)

(51) Int.Cl.⁷

識別記号

F I

テマート(参考)

G 0 6 F 9/46

3 5 0

G 0 6 F 9/46

3 5 0 5 B 0 6 5

3/06

3 0 1

3/06

3 0 1 J 5 B 0 9 8

3 0 1 K

審査請求 未請求 請求項の数 8 O L (全 14 頁)

(21) 出願番号 特願2002-120216(P2002-120216)

(22) 出願日 平成14年4月23日 (2002.4.23)

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72) 発明者 長須賀 弘文

神奈川県川崎市麻生区王禅寺1099番地 株

式会社日立製作所システム開発研究所内

(72) 発明者 清井 雅広

神奈川県川崎市麻生区王禅寺1099番地 株

式会社日立製作所システム開発研究所内

(74) 代理人 100075096

弁理士 作田 康夫

最終頁に続く

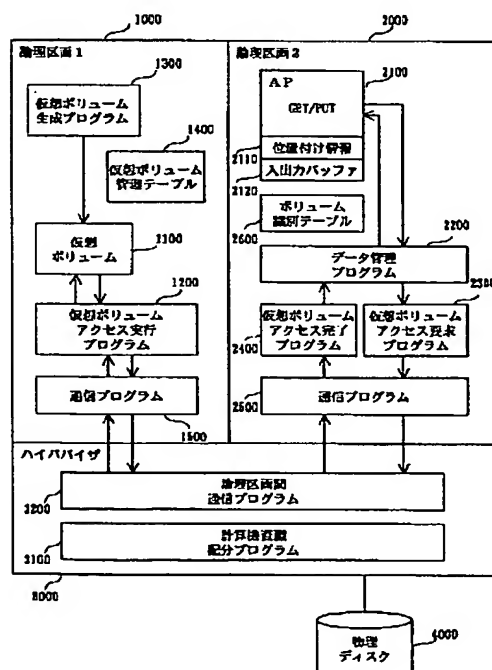
(54) 【発明の名称】 実記憶利用方法

(57) 【要約】 (修正有)

【課題】 外部記憶装置に共用データを配置しているシステムでは、アクセスの集中による遅延の発生や、この遅延発生を回避する手法を採用するために発生するユーザプログラムの変更の必要性といった課題があった。

【解決手段】 仮想計算機システムにおいて、選択した任意の計算機には、他の計算機から、通信手段によって送られた外部記憶装置への入出力要求コマンドの中で指定されたアドレスを、当該選択された論理区画が保持する実記憶装置内の対応するデータ域のアドレスに変換するステップを設ける。そして、そのコマンドが入力要求の時は、変換したアドレスのデータ域に格納されているデータをコマンドの送信元である計算機へ返送するステップを設ける。さらに、コマンドが出力要求の時は、コマンドと一緒に送られてきたデータを、変換したアドレスのデータ域に格納するステップを設ける。

図 1



【特許請求の範囲】

【請求項 1】 物理的記憶装置を論理的に分割した論理区画ごとに記憶されたプログラムを実行させることにより、複数の計算機を仮想的に実行することができる仮想計算機システムにおいて、前記論理区画のうち少なくとも 1 つの第 1 の論理区画に記憶されたプログラムの実行を選択する選択手段とを有し、少なくとも前記第 1 の論理区画には、第 2 の論理区画との間で通信可能な通信手段と、前記物理的記憶装置に対するデータの入力要求コマンドを受信する受信手段と、前記入力要求コマンドにおいて指定されるデータアドレスを、前記第 1 の論理区画に割当てられた論理的データアドレスに変換する変換手段と、前記論理的データアドレスに基づいて、前記物理的記憶装置に対して前記データを格納する格納手段とを有することを特徴とする仮想計算機システム。

【請求項 2】 請求項 1 に記載の仮想計算機システムにおいて、少なくとも前記第 1 の論理区画には、さらに、前記選択手段により選択された第 1 の論理区画に記憶されたプログラムを実行することにより発行された、前記第 2 の論理区画に対するデータの出力要求に基づいて、前記物理的記憶装置に対する出力要求コマンドを生成する生成手段とを有し、前記通信手段は、前記入力要求コマンドと前記出力要求コマンドを用いて第 2 の論理区画に対して格納するデータとを、前記第 2 の論理区画に対して送信することを特徴とする仮想計算機システム。

【請求項 3】 物理的記憶装置を論理的に分割した論理区画ごとに記憶されたプログラムを実行させることにより、複数の計算機を仮想的に実行することができる仮想計算機システムにおいて、前記論理区画のうち少なくとも 1 つの第 1 の論理区画に記憶されたプログラムの実行を選択する選択手段とを有し、少なくとも前記第 1 の論理区画には、前記第 2 の論理区画との間で通信可能な通信手段と、前記物理的記憶装置に対するデータの出力要求コマンドを受信する受信手段と、前記出力要求コマンドにおいて指定されるデータアドレスを、前記第 1 の論理区画に割当てられた論理的データアドレスに変換する変換手段とを有し、前記通信手段は、前記論理的データアドレスに基づいて、前記物理的記憶装置に格納されたデータを、前記入力要求コマンドを発行した第 2 の論理区画に対して送信することを特徴とする仮想計算機システム。

【請求項 4】 請求項 3 に記載の仮想計算機システムにおいて、

少なくとも前記第 1 の論理区画には、さらに、前記選択手段により選択された第 1 の論理区画に記憶されたプログラムを実行することにより発行された、第 2 の論理区画に対するデータの出力要求に基づいて、前記物理的記憶装置に対する出力要求コマンドを生成する生成手段とを有し、前記通信手段は、前記出力要求コマンドを前記第 2 の論理区画の通信手段に対して送信することを特徴とする仮想計算機システム。

【請求項 5】 物理的記憶装置を論理的に分割した論理区画ごとに記憶されたプログラムを実行させることにより、複数の計算機を仮想的に実行することができる仮想計算機システムによる実記憶利用方法において、前記論理区画のうち少なくとも 1 つの第 1 の論理区画に記憶されたプログラムの実行を選択する選択ステップと、

少なくとも前記第 1 の論理区画において、第 2 の論理区画との間で通信可能な通信ステップと、前記物理的記憶装置に対するデータの出力要求コマンドを受信する受信ステップと、前記入力要求コマンドにおいて指定されるデータアドレスを、前記第 1 の論理区画に割当てられた論理的データアドレスに変換する変換ステップと、前記論理的データアドレスに基づいて、前記物理的記憶装置に対して前記データを格納する格納ステップとを有することを特徴とする実記憶利用方法。

【請求項 6】 物理的記憶装置を論理的に分割した論理区画ごとに記憶されたプログラムを実行させることにより、複数の計算機を仮想的に実行することができる仮想計算機システムによる実記憶利用方法において、前記論理区画のうち少なくとも 1 つの第 1 の論理区画に記憶されたプログラムの実行を選択する選択ステップとを有し、

少なくとも前記第 1 の論理区画において、前記第 2 の論理区画との間で通信可能な通信ステップと、前記物理的記憶装置に対するデータの出力要求コマンドを受信する受信ステップと、前記出力要求コマンドにおいて指定されるデータアドレスを、前記第 1 の論理区画に割当てられた論理的データアドレスに変換する変換ステップとを有し、前記通信ステップは、前記論理的データアドレスに基づいて、前記物理的記憶装置に格納されたデータを、前記入力要求コマンドを発行した第 2 の論理区画に対して送信することを特徴とする実記憶利用方法。

【請求項 7】 複数の計算機ノードと、該複数の計算機ノ

ードが共有する物理記憶装置とを有する計算機システムにおいて、
前記のうち少なくとも1つの第1の計算機ノードを選択する選択手段とを有し、
少なくとも前記第1の計算機ノードには、
第2の計算機ノードとの間で通信可能な通信手段と、
前記物理的記憶装置に対するデータの入力要求コマンドを受信する受信手段と、
前記入力要求コマンドにおいて指定されるデータアドレスを、前記第1の論理区画に割当てられた論理的データアドレスに変換する変換手段と、
前記論理的データアドレスに基づいて、前記物理的記憶装置に対して前記データを格納する格納手段とを有することを特徴とする計算機システム。

【請求項8】複数の計算機ノードと、該複数の計算機ノードが共有する物理記憶装置とを有する計算機システムにおいて、
前記のうち少なくとも1つの第1の計算機ノードを選択する選択手段とを有し、
少なくとも前記第1の計算機ノードは、
前記第2の計算機ノードとの間で通信可能な通信手段と、
前記物理的記憶装置に対するデータの出力要求コマンドを受信する受信手段と、
前記出力要求コマンドにおいて指定されるデータアドレスを、前記第1の計算機ノードに割当てられた論理的データアドレスに変換する変換手段とを有し、
前記通信手段は、
前記論理的データアドレスに基づいて、前記物理的記憶装置に格納されたデータを、前記入力要求コマンドを発行した第2の計算機ノードに対して送信することとを特徴とする計算機システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】大容量の実記憶装置の利用方法に関し、特に複数の計算機を仮想的に実行することができる仮想計算機システムにおける実記憶装置の利用方法に関する。

【0002】

【従来の技術】近年の基幹系のサーバシステムでは、膨大するデータ量や処理量に柔軟に対応することを可能にするため、複数の計算機イメージから構成されるクラスタ型のシステム構成を採用するのが一般的である。さらに、多くの計算機システムでは、各計算機イメージ間で共用が必要なデータは、各計算機イメージからアクセスできる共用の外部記憶装置に配置している。

【0003】しかし、共用データに対しては、複数の計算機システムからのアクセス要求が発生するため、データの整合性を保証するため、排他処理を実施しながらデータアクセスを制御する必要がある。一般的には、こうし

た排他処理用に利用する制御情報も、共用の外部記憶装置に配置する。そのため、システムの応答性能やスループットの向上を図るためには、外部記憶装置に配置した排他処理用のデータに高速にアクセスできる環境が必要である。

【0004】しかし、外部記憶装置内では、ディスクを回転させながら、対象となる領域へアクセスするため、共用の外部記憶装置に対してアクセスが集中した場合は、ディスクの回転待ちが、アクセスの遅延を発生させる要因となる。

【0005】それを解決するための手段として特開平9-293055がある。特開平9-293055では、共用の外部記憶装置の代わりに拡張メモリを配置し、このメモリ域に共用ファイルを配置すると共に、排他処理で利用するロック領域を設けている。こうすることで、共用するデータへのアクセス時の回転待ち等によるアクセス遅延を回避すると共に、制御情報へのアクセスの高速化により、排他処理の効率化を図っている。

【0006】

【発明が解決しようとする課題】しかし、上記従来技術を用いるためには、新たにアプリケーションプログラムを用意する必要性が生ずる。共用するデータを共用外部記憶装置に配置していたシステムでは、外部記憶装置へアクセスするようプログラムを作成していた。しかし、上記従来技術を利用するためには、それを、共用メモリへアクセスするように変更する必要がある。

【0007】また、上述した課題を解決するための手法として、オペレーティングシステムが用意している外部記憶装置への入出力用インタフェースを用いて、上述した共用メモリにアクセスできる機能をオペレーティングシステムがサポートする手法を想定することができる。この手法では、外部記憶装置へアクセスするためにアプリケーションプログラムが発行した入出力要求を、オペレーティングシステムがトラップし、上記の共用メモリへのアクセスに変更する。そのため、各計算機ノードで動作しているオペレーティングシステムは、上記共用メモリに、外部記憶装置への入出力インタフェースを用いて、アクセスする機能をサポートする必要がある。オペレーティングシステムが、こうした機能をサポートするためには、膨大な開発量を要する。

【0008】本発明の第1の目的は、アプリケーションプログラムに対しては、従来通りの外部記憶装置へアクセスするインタフェースでの動作を保証した上で、比較的少量のオペレーティングシステムの開発量で、共用データへのアクセスの高速化によるシステムの応答性能やスループットの向上を図る方法を提供することである。

【0009】本発明の第2の目的は、こうしたデータをへのアクセスに関して、別の論理区画域に配置させることで、データの機密性を向上させることである。

【0010】

【課題を解決するための手段】上記課題を解決するために、論理分割により、ひとつの計算機システム上に複数の計算機イメージを構築することが可能なシステムにおいて、選択した任意の計算機イメージには、他の計算機イメージから、通信手段によって送られてきた外部記憶装置への入出力要求コマンド内で指定されたアドレスを、当該選択された論理区画が保持する実記憶装置内の対応するデータ域のアドレスに変換するステップを設ける。さらに、入出力要求コマンドが入力要求の時は、変換するステップにより求められたアドレスのデータ域に格納されているデータを入出力要求コマンドの送信元である計算機イメージへ返送し、入出力要求コマンドが出力要求の時は、入出力要求コマンドと一緒に送られてきたデータを、変換するステップにより求められたアドレスのデータ域に格納するステップを設ける。

【0011】また、選択された計算機イメージ以外には、動作中のプログラムが発行したデータの入出力要求から、外部記憶装置への入出力要求コマンドを生成するステップを設ける。さらに、この入出力要求コマンドが、上記選択された計算機イメージのいずれかに格納されたデータの入力要求である場合には、当該入出力要求コマンドを、アクセス対象のデータが格納されている計算機イメージへ通信によって送り、入出力要求コマンドが選択された計算機イメージのいずれかへの出力要求である場合には、格納先の計算機イメージに、この入出力要求コマンドとデータを通信によって送るステップを設ける。

【0012】なお、この解決手段は、互いに通信することが可能な複数の計算機ノードから構成されるクラスタ型計算機システムにおいても、適用することができる。

【0013】

【発明の実施の形態】図1は、本発明の実施の形態を示す計算機システムの構成を示す。図1において、1000、2000はそれぞれ論理区画1、論理区画2である。また、3000は、計算機システム内の資源を各論理区画に割当て、また、論理区画間で必要な通信を行うハイパバイザである。ハイパバイザ(3000)によって、計算機システム内の資源を割当てられた論理区画1(1000)及び論理区画2(2000)に記憶されたプログラムを制御部(CPU)により実行させることにより、仮想的に独立した複数の計算機として実行することができる。

【0014】4000は、計算機システムに接続された物理ディスク装置であり、本実施の形態では、ハイパバイザ(3000)により、論理区画2(2000)に割当てられている。

【0015】1100は、論理区画1(1000)に割当てられた実記憶装置内に生成された仮想ボリュームである。この仮想ボリューム(1100)を生成するのが、仮想ボリューム生成プログラム(1200)であ

る。

【0016】1300は、仮想ボリュームアクセス実行プログラムであり、論理区画2(2000)から仮想ボリューム(1100)に対するアクセス要求を受け取ったとき、その要求内容に応じて、仮想ボリューム(1100)領域へのデータ書き込み、データ読み込み、または処理対象データ位置の計算を行う。

【0017】1400は、仮想ボリューム管理テーブルであり、仮想ボリューム(1100)を管理するために必要な各種情報を格納する。

【0018】1500は、他の論理区画あるいはネットワークを介した外部の計算機との通信処理を実行する通信プログラムである。通信プログラム(1500)は、論理区画1(1000)から他の計算機に向けて通信を発信する処理と、他の計算機から論理区画1(1000)への通信を受け付け、通信内容に応じて所定のプログラムを実行する処理を有する。通信内容が、仮想ボリューム(1100)に対するアクセス要求である場合は、前述した仮想ボリュームアクセス実行プログラム(1200)を起動する。また、仮想ボリュームアクセス処理が完了したあとに、その応答を要求元の論理区画1(1000)へ送信するのも、通信プログラム(1500)である。

【0019】続いて、論理区画2の構成内容を説明する。2100は論理区画2(2000)内で実行されるアプリケーションプログラム(以下AP)である。AP(2100)は、物理ディスク(4000)、あるいは仮想ボリューム(1100)への入出力処理の実行要求を含む。AP(2100)内の2110は、入出力処理を実行する際、物理ディスク(4000)あるいは仮想ボリューム(1100)内での処理対象場所を示す位置付け情報である。また、2120は、入出力処理を行うデータを格納する入出力バッファである。

【0020】2200は、AP(2100)からのデータ入出力要求を受け、実際の入出力処理を実行するデータ管理プログラムである。

【0021】2300は、仮想ボリュームアクセス要求プログラムである。仮想ボリュームアクセス要求プログラム(2300)は、AP(2100)から要求されたデータ入出力の処理対象が仮想ボリュームであったときに、データ管理プログラム(2200)によって呼び出される。そして、仮想ボリュームの管理元である論理区画1(1000)に対する通信の実行を、通信プログラム(2500)に要求する。

【0022】2400は、仮想ボリュームアクセス完了プログラムである。仮想ボリューム(1100)に対するアクセスが完了すると、仮想ボリューム(1100)の管理元である論理区画1(1000)から、処理完了を示す応答が送信される。論理区画2(2000)の通信プログラム(2500)は、このアクセス完了の通信

を受け取ると、仮想ボリュームアクセス完了プログラム（2400）を実行する。AP（2100）からの入出力要求がデータ読み込み要求の場合、仮想ボリュームアクセス完了プログラム（2400）は、仮想ボリューム（1100）から読み込んだデータを入出力バッファ領域に複写する。

【0023】通信プログラム（2500）は、論理区画1（1000）内の通信プログラム（1500）と同様に、他の論理区画あるいはネットワークを介した外部の計算機との通信処理を実行する。

【0024】続いて、ハイパバイザ（3000）の構成内容を説明する。3100は、計算機資源配分プログラムである。計算機資源配分プログラム（3100）は、実記憶装置、プロセッサといった計算機が有する資源を、論理区画1（1000）及び論理区画2（2000）に静的または動的に割当てて。これにより、論理区画1（1000）、論理区画2（2000）が、それぞれ独立した計算機システムイメージとして稼働できる環境を提供する。

【0025】3200は、ハイパバイザ（3000）内の論理区画間通信プログラムである。論理区画間通信プログラム（3200）は、論理区画1（1000）から論理区画2（2000）へ、またはその逆方向の通信要求がなされたとき、要求先の論理区画に通信データを伝達する。

【0026】図2は、AP（2100）から仮想ボリューム（1100）に対する入出力要求が発生したときの、制御の流れを示す。データ管理プログラム（2200）は、仮想ボリュームアクセス要求プログラム（2300）を起動し、その後、通信プログラム（2500）によって、論理区画1（1000）へアクセス要求が発信される。この通信は、ハイパバイザ（3000）の論理区画間通信プログラム（3200）を経由して、論理区画1（1000）内の通信プログラム（1500）へ渡される。

【0027】要求を受け取った論理区画1（1000）側では、仮想ボリュームアクセス実行プログラム（1200）を起動し、要求内容に従い仮想ボリューム（1100）へのデータ書き込み、あるいはデータ読み込みを行ったのち、要求元の論理区画2（2000）へ応答を返す。

【0028】応答情報を受け取った論理区画2（2000）内の通信プログラム（2500）は、仮想ボリュームアクセス完了プログラム（2400）を起動し、以下、データ管理プログラム（2200）を介して、AP（2100）の処理を再開する。

【0029】続いて、図3を用いて、仮想ボリューム管理テーブル（1400）の構成を説明する。仮想ボリューム管理テーブル（1400）は、仮想ボリューム（1100）を、物理ディスク（4000）と同じインタフ

ェースで利用するための構成情報と、各時点における仮想ボリューム（1100）内でのアクセス場所を管理するための情報とから構成される。ここで、物理ディスク（4000）内のデータをアクセスする場合、物理ディスク（4000）内におけるデータの配置場所は、シリンダ番号、ヘッダ番号、及び、シリンダ番号とヘッダ番号とにより決定されるトラック内レコード番号によって特定される。各ボリュームのシリンダ数、ヘッダ数、トラック長は、装置毎に物理的に決められる。

【0030】このような物理ディスク（4000）へのアクセスと共通のインタフェースに対応するため、仮想ボリューム管理テーブル（1400）は、仮想シリンダ数（1410）、仮想ヘッダ数（1420）、仮想トラック長（1430）を格納するフィールドを有する。これらのフィールドに格納された値により、仮想ボリューム（1100）を構成するデータ領域の大きさが決定される。

【0031】また、仮想ボリューム管理テーブル（1400）内のレコード長（1440）は、AP（2100）からアクセスするデータの最小単位を表すレコード長を格納する領域である。更に、データ先頭ポインタ（1450）は、仮想ボリューム（1100）の領域を確保した実記憶装置内での先頭アドレスを格納するフィールドである。アクセスポインタ（1460）は、各時点においてアクセスすべき仮想ボリューム（1100）内のデータ位置を示すアドレスを格納するフィールドである。

【0032】仮想ボリューム管理テーブル（1400）を構成する各フィールドの初期設定と、仮想ボリューム（1100）データ領域の確保を行うのが、仮想ボリューム生成プログラム（1300）である。図4は、仮想ボリューム生成部（1300）が、仮想ボリューム管理テーブル（1400）を初期設定した状態を表した図である。ここで、仮想シリンダ数（1410）が1024個、仮想ヘッダ数（1420）が16個、仮想トラック長（1430）が32768バイトである仮想ボリューム（1100）が生成されている。また、各トラックに格納するデータ入出力の単位であるレコード長（1440）の初期値として、4096バイトを設定している。

【0033】本実施の形態では、仮想ボリューム（1100）を構成するデータ領域は、論理区画1（1000）の実記憶装置内に、連続して確保される。データ配置順序は、シリンダ番号の昇順、及び各シリンダ内のトラック番号の昇順である。ただし、シリンダ番号、トラック番号とも0から開始される。図4における仮想ボリューム（1100）データ領域の説明で、「シリンダ*i*のトラック*j*」とは、「シリンダ番号が*i*で、かつ、ヘッダ番号が*j*であるトラック」に対応するデータ領域であることを意味する。

【0034】図4で、仮想ボリューム管理テーブル（1

400)内のデータ先頭ポインタ(1450)の値は4096に設定されている。すなわち、本実施の形態では、仮想ボリューム(1100)用のデータ領域を確保した先頭番地は、論理区画1(1000)の実記憶装置上で4096番地であることを表す。

【0035】図4で、アクセスポインタ(1460)の値は528384に設定されている。すなわち、次に仮想ボリューム(1100)に対するアクセス要求が発生したときに、論理区画1(1000)の実記憶装置上で528384番地に位置するデータを処理することを表す。ここで、528384番地とは、「データ先頭ポインタ(2450)+仮想ヘッダ数(2420)×仮想トラック長(2430)」と一致し、仮想ボリューム(1100)内のシリンダ1、トラック0に相当する。

【0036】図5は、論理区画2(2000)内のボリューム識別テーブル(2600)の構成を示している。ボリューム識別テーブル(2600)は、論理区画2(2000)内で実行されるAP(2100)がアクセス可能なディスク装置の一覧を、ボリュームID(2610)とボリューム種別(2620)の組合せにより示している。この情報は、システム構成情報として、システム立上げ時のパラメタ等によって、利用者があらかじめ登録した情報を格納する。

【0037】ボリュームID(2610)は、アクセス対象のディスクを識別するために利用する情報である。本実施の形態では、登録順に1、2、...なる昇順の整数により、各ディスクを識別する。システムに登録された順序は、物理ディスク(4000)、仮想ボリューム(1100)であるとする。すなわち、ボリュームID(2610)=1は物理ディスク(4000)に、また、ボリュームID(2610)=2は仮想ボリューム(1100)に、それぞれ対応している。

【0038】ボリューム種別(2620)は、ボリュームID(2610)で示されたディスク装置の種類が、物理ディスクか、あるいは仮想ボリュームであるかを識別するための情報を格納するフィールドである。本実施の形態では、ボリューム種別(2620)=Pは物理ディスク、ボリューム種別(2620)=Vは仮想ボリュームであることを表す。

【0039】図6は、論理区画2(2000)内のAP(2100)から要求された入出力動作を実行するために必要な情報を格納したチャンネルコマンド語(以下CCW)の構成を示している。5000は、物理ディスク(4000)に対して処理を要求するときに生成するチャンネルコマンド語の構成を示している。CCW(5000)は、データ管理プログラム(2200)が生成する。また、5100は、仮想ボリューム(1100)に対するアクセス要求時、論理区画1(1000)との間の通信に利用される通信CCWの構成を示している。通信CCW(5100)は、データ管理プログラム(22

00)によって起動された仮想ボリュームアクセス要求プログラム(2300)が生成する。

【0040】CCW(5000)は、コマンド(5010)とデータアドレス(5020)を有する。コマンド(5010)は、入出力装置に対してどのような処理を要求するかを示す領域で、本実施の形態では、位置付け、書込み、読込みの三種類を指定できる。このうち、位置付けコマンドとは、ディスク装置内での位置情報(位置付け情報)を指定し、次に処理すべきデータ位置を指示するコマンドである。

【0041】データアドレス(5020)は、処理対象データが存在する実記憶装置上でのアドレスを示す領域である。位置付け要求の場合は、前記した位置付け情報が存在するアドレスが、このデータアドレス(5020)で指定される。

【0042】対象が仮想ボリューム(1100)の場合は、CCW(5000)をもとにして、仮想ボリュームアクセス要求プログラム(2300)が通信CCW(5100)を生成する。更に、仮想ボリューム(1100)アクセス完了後、仮想ボリュームアクセス実行プログラム(1200)からの応答時にも、この通信CCW(5100)が生成される。通信CCW(5100)は、コマンド(5110)、データアドレス(5120)、データ(5130)によって構成される。このうち、コマンド(5110)、データアドレス(5120)は、それぞれCCW(5000)のコマンド(5010)、データアドレス(5020)と同じ意味を持つ。データ(5130)は、コマンド(5110)によって示された要求種別に従い、データ本体も転送する必要がある場合に、そのデータ本体を格納するための領域である。

【0043】以下、仮想ボリューム(1100)をアクセスするときのCCW(5000)と通信CCW(5100)の生成状況を、位置付け要求、書込み要求、読込み要求の順に説明する。

【0044】図7は、AP(2100)から仮想ボリューム(1100)に対して位置付け要求を行う際の論理区画2(2000)側の状態を示す。CCW(5000)のデータアドレス(5020)に、位置付け情報(2110)のアドレスが格納される。この位置付け情報(2110)は、論理区画2(2000)の実記憶装置内に存在する。そのため、論理区画1(1000)から直接、内容を読み取ることはできない。そこで、論理区画1(1000)に対する通信CCW(5100)のデータ(5130)に、データアドレス(5020)が示す位置付け情報(6000)を格納し、通信CCW(5100)を論理区画1(1000)に送信する。

【0045】図8は、位置付け要求通信を受け取った論理区画1(1000)側の状態を示す。論理区画1(1000)側では、通信CCW(5100)内データ(5

130)に格納された位置付け情報をもとにして、仮想ボリューム管理テーブル(1400)内のアクセスポインタ(1460)を更新する。元の位置付け情報は、物理ディスク装置内での配置場所を記述した形式となっているため、この情報を論理区画1(1000)内で仮想ボリューム(1100)を配置した実記憶装置内でのアドレスに変換する必要がある。この変換処理の詳細については、のちほど仮想ボリュームアクセス実行プログラム(1200)の処理フローチャートを説明する際に、詳しく説明する。

【0046】位置付け処理が完了したのち、論理区画2(2000)に対する応答として、コマンド(5110)、データアドレス(5120)、データ(5130)の内容が要求時と同じ通信CCW(5100)を生成し、論理区画2(2000)に通信する。

【0047】引き続き、図9、図10を用いて、仮想ボリューム(1100)への書き込み処理を行う際の状態を説明する。図8は、データの書き込みを要求する論理区画2(2000)側の状態を示す。CCW(5000)内のデータアドレス(5020)は、AP(2100)から仮想ボリュームへ書き込むべきデータが格納された入出力バッファ(2120)のアドレスを示している。この入出力バッファ(2120)の内容をデータ(5130)に格納した通信CCW(5100)を生成し、論理区画1(1000)に送信する。

【0048】図10は、書き込み要求を受け取った論理区画1(1000)側の状態を示している。通信CCW(5100)内データ(5130)に格納された内容を、仮想ボリューム(1100)のうち、仮想ボリューム管理テーブル(1400)内のアクセスポインタ(1460)によって示されるデータ領域(1110)に複写する。論理区画2(2000)に対する応答の通信CCW(5100)では、データ(5130)に情報を格納する必要はない。

【0049】引き続き、図11から図13を用いて、仮想ボリューム(1100)への読み込み要求を行う際の状態を説明する。

【0050】図11は、読み込みを要求する論理区画2(2000)側の状態を示す。論理区画1(1000)に対して送信する通信CCW(5100)のうち、コマンド(5000)、データアドレス(5100)の内容はCCW(5000)と同様である。データ(5130)には、アクセス要求時には情報を格納する必要はない。

【0051】図12は、読み込み要求通信を受け取った論理区画1(1000)側の状態を示している。アクセスポインタ(1460)が示す仮想ボリューム(1100)内のデータ領域(1110)から、通信CCW(5100)内データ(5130)へデータを複写して、論理区画2(2000)に通信CCW(5100)を応答

する。

【0052】図13は、読み込み応答通信を受け取った論理区画2(2000)側の状態を示している。通信CCW(5100)内データ(5130)に格納された内容、すなわち仮想ボリューム(1100)から読み込んだデータを、データアドレス(5120)が示すAP(2100)内の入出力バッファ(2120)に複写する。

【0053】図14は、仮想ボリューム(1100)へのアクセス要求を受け取ったときに実行される論理区画1(1000)内の仮想ボリュームアクセス実行プログラム(1200)の処理フローチャートである。

【0054】まず、応答通信の準備として、要求された通信CCW(5100)内のコマンド(5110)、データアドレス(5120)を、応答通信用の通信CCW(5100)のコマンド(5110)、データアドレス(5120)に格納(ステップ1210)する。

【0055】次に、要求されたコマンド種別を判定(ステップ1220)する。ここで、コマンド種別には、位置付け、書き込み、読み込みの三種類が存在する。コマンド種別が位置付け要求だった場合は、通信CCW(5100)内データ(5200)に格納されている位置付け情報を、仮想ボリューム(1100)を格納した論理区画1(1000)内の実記憶装置のアドレスに変換(ステップ1230)する。位置付け情報として指示される指定は、物理ディスク(4000)をアクセスするための情報と共通であり、「CCHHRR」という形式で示される。ここで、CCはシリンダ番号、HHはヘッダ番号、Rはレコード番号である。このような指定に対し、仮想ボリューム(1100)内のデータ配置先として、「データ先頭ポインタ(2450)+(CC×仮想ヘッダ数(2420)+HH)×仮想トラック長(2430)+R×レコード長(2440)」なる計算を実行することによって、仮想ボリューム(1100)内データ領域のアドレスに変換する。更に、変換したデータを、アクセスポインタ(1460)に格納(ステップ1240)する。

【0056】コマンド種別が書き込み要求だった場合は、通信CCW(5100)内データ(5130)の内容を、アクセスポインタ(1460)が示す仮想ボリューム(1100)内のデータ領域に複写(ステップ1250)する。また、コマンド種別が読み込み要求だった場合は、アクセスポインタ(1460)が示す仮想ボリューム(1100)内のデータ領域から、応答通信用の通信CCW(5100)内データ(5130)へ内容を複写(ステップ1260)する。

【0057】最後に、以上のステップで作成した応答通信用の通信CCW(5100)を、要求元の論理区画2(2000)に対して送信することを、通信プログラム(1500)に依頼(ステップ1270)する。

【0058】図15は、論理区画2（2000）において、AP（2100）から入出力要求を受け取ったときに実行されるデータ管理プログラム（2200）の処理フローチャートである。

【0059】はじめに、入出力装置に対して要求を実行するためのCCWを作成（ステップ2210）する。次に、入出力要求を実行する装置のボリュームIDを取得（ステップ2220）する。引き続き、アクセス対象入出力装置が、仮想ボリュームか、あるいは物理ディスクかを判定（ステップ2230）する。この判定処理は、ボリューム識別テーブル（2600）を参照し、ステップ2220で取得したボリュームID（2610）に対応するボリューム種別（2620）をチェックすることによって実現できる。

【0060】アクセス対象が、物理ディスクであった場合、該当する入出力装置に対して入出力起動命令を実行（ステップ2240）する。この物理ディスク装置へアクセスする場合の処理は、論理区画1（1000）内に仮想ボリューム（1100）を生成しない計算機システムと同様である。

【0061】ステップ2230の判定において、アクセス対象が仮想ボリューム（1100）だった場合、仮想ボリューム（1100）が存在する論理区画1（1000）への通信を実行するため、仮想ボリュームアクセス要求プログラム（2300）を実行（ステップ2250）する。

【0062】図16は、仮想ボリュームアクセス要求プログラム（2300）の処理フローチャートである。

【0063】通信CCW（5100）を構成するコマンド（5110）、データアドレス（5120）に、それぞれデータ管理プログラム（2200）が作成したCCW（5000）内のコマンド（5010）、データアドレス（5020）と同様の情報を格納（ステップ2310）する。次に、コマンド種別を判定（ステップ2320）する。判定の結果、コマンド種別が位置付け、または書込みだった場合、通信CCW（5100）のデータ（5130）に、データアドレス（5120）が示す領域の内容を格納（ステップ2330）する。この情報は、位置付け要求の場合は位置付け情報（6000）であり、書込み要求の場合は仮想ボリューム（1100）に書き込むべき入出力バッファ（7000）内のデータである。

【0064】以上の処理ステップによって作成した通信CCW（5100）を、仮想ボリューム（1100）が存在する論理区画1（1000）に対して送信することを、通信プログラム（2500）に依頼（ステップ2340）する。

【0065】図17は、仮想ボリューム（1100）へのアクセスが完了したとき、論理区画1（1000）から応答された通信CCW（5100）を受け取った通信

プログラム（2500）が起動する仮想ボリュームアクセス完了プログラム（2400）の処理フローチャートである。

【0066】まず通信CCW（5100）の内容から、コマンド種別を判定（ステップ2410）する。同判定において、コマンド種別が読み込みだった場合、通信CCW（5100）内データ（5130）の内容を、データアドレス（5120）が示す実記憶装置内の領域、すなわちAP（2100）が指定した入出力バッファ（7000）へ複写（2420）する。ステップ2410の判定において、コマンド種別が位置付け、または書込みであった場合は、通信CCW（5100）からのデータ複写は必要ない。

【0067】続けて、入出力要求元のAP（2100）を再起動するために、データ管理プログラム（2200）を実行（ステップ2430）する。

【0068】次に、図18を用いて、本発明を用いた第2の実施の形態を説明する。

【0069】第2の実施の形態では、論理区画1（1000）、論理区画2（2000）に加え、論理区画3（6000）と、論理区画3（6000）に割当てられた物理ディスク（7000）が存在する。論理区画3（6000）の構成内容は論理区画2（2000）と同様であり、説明は省略する。このような構成をとることによって、論理区画2（2000）で実行されるAP（2100）と、論理区画3（6000）で実行されるAP（6100）が、仮想ボリューム（1100）内のデータを共用することができる。

【0070】次に、図19を用いて、本発明を用いた第3の実施の形態を説明する。

【0071】第3の実施の形態における計算機システムは、1つの計算機を論理区画に分割して実行するシステムではなく、複数の計算機ノードから成るクラスタ型計算機システムである。ここで、計算機システムは、計算機ノード1（8000）と、計算機ノード2（9000）とから構成される。また、10000は、計算機ノード2（9000）に接続された物理ディスク装置であり、11000は、計算機ノード1（8000）と計算機ノード2（9000）の間でデータ通信を実行するための通信経路である。

【0072】計算機ノード1（8000）内の構成内容は、第1の実施の形態における論理区画1（1000）と同様である。同じように、計算機ノード2（9000）内の構成内容は、第1の実施の形態における論理区画2（2000）と同様である。仮想ボリューム（8100）をアクセスするために、計算機ノード間で通信するデータの形式も、第1の実施の形態で説明した通信CCW（5100）と同じ構成とする。これにより、クラスタ型計算機システムにおいても、計算機ノード1（8000）の実記憶装置上に生成した仮想ボリューム（8

100)を、計算機ノード2(9000)のAP(9100)から利用することが可能となる。更に計算機ノード数を増加した構成では、仮想ボリューム(8100)内のデータを、複数の計算機ノード間で共用することができる。

【0073】

【発明の効果】本発明によれば、共用するデータを外部記憶装置に配置していたシステムで動作していたプログラムを変更せずに、共用データへのアクセスの高速化を図ることができるので、システムの応答性能やスループットを向上できる。また、本発明によれば、共用しないデータに関しても、メモリを利用することで、アクセスの高速化を図り、システムの応答性能やスループットを向上できる。また、本発明によれば、こうしたデータへのアクセスに関して、別の論理区画域に配置させることで、データの機密性を向上させることができる。

【図面の簡単な説明】

【図1】 第1の実施の形態を示す計算機システムの構成図

【図2】 第1の実施の形態における制御の流れ

【図3】 第1の実施の形態における仮想ボリューム管理テーブルの構成図

【図4】 第1の実施の形態における仮想ボリューム初期設定の状態

【図5】 第1の実施の形態におけるボリューム識別テーブルの構成図

【図6】 第1の実施の形態におけるCCWと通信CCWの構成図

【図7】 第1の実施の形態における位置付け要求時の論理区画2の状態

【図8】 第1の実施の形態における位置付け要求時の論理区画1の状態

【図9】 第1の実施の形態における書き込み要求時の論理区画2の状態

【図10】 第1の実施の形態における書き込み要求時の論理区画1の状態

【図11】 第1の実施の形態における読み込み要求時の論理区画2の状態

【図12】 第1の実施の形態における読み込み要求時の論理区画1の状態

【図13】 第1の実施の形態における読み込み要求時の論理区画2の状態

【図14】 第1の実施の形態における仮想ボリュームアクセス実行プログラムの処理フローチャート

【図15】 第1の実施の形態におけるデータ管理プログラムの処理フローチャート

【図16】 第1の実施の形態における仮想ボリュームアクセス要求プログラムの処理フローチャート

【図17】 第1の実施の形態における仮想ボリュームアクセス完了プログラムの処理フローチャート

【図18】 第2の実施の形態を示す計算機システムの構成図

【図19】 第3の実施の形態を示す計算機システムの構成図

【符号の説明】

1000……論理区画1、1100……仮想ボリューム、1200……仮想ボリュームアクセス実行プログラム、1300……仮想ボリューム生成プログラム、1400……仮想ボリューム制御テーブル、1500……通信プログラム、2000……論理区画2、2100……アプリケーションプログラム、2110……位置付け情報、2120……入出力バッファ、2200……データ管理プログラム、2300……仮想ボリュームアクセス要求プログラム、2400……仮想ボリュームアクセス完了プログラム、2500……通信プログラム、3000……ハイパバイザ、3100……計算機資源配分プログラム、3200……論理区画間通信プログラム、4000……物理ディスク

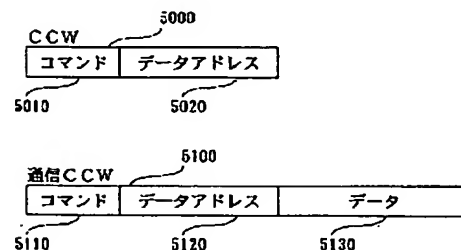
【図5】

図5

ボリューム識別テーブル	
ボリュームID	ボリューム種別
1	P
2	V

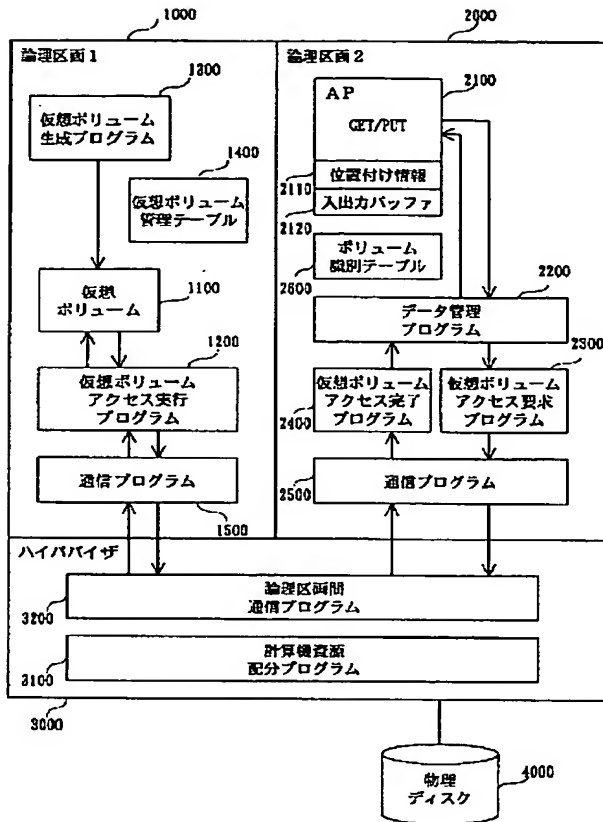
【図6】

図6



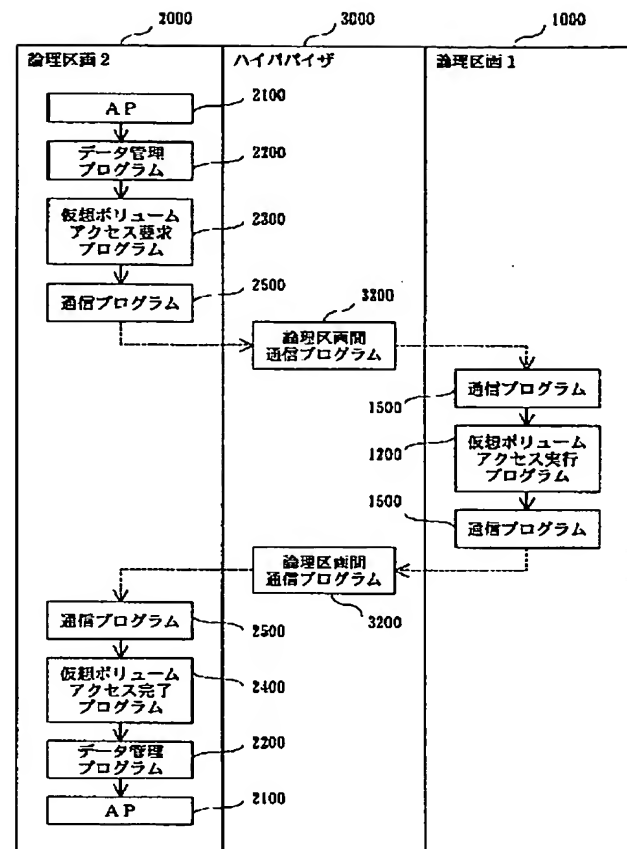
【図1】

図1



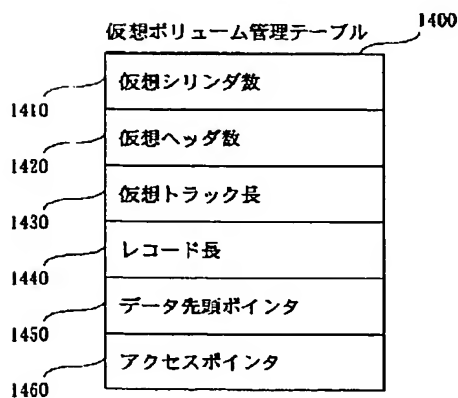
【図2】

図2



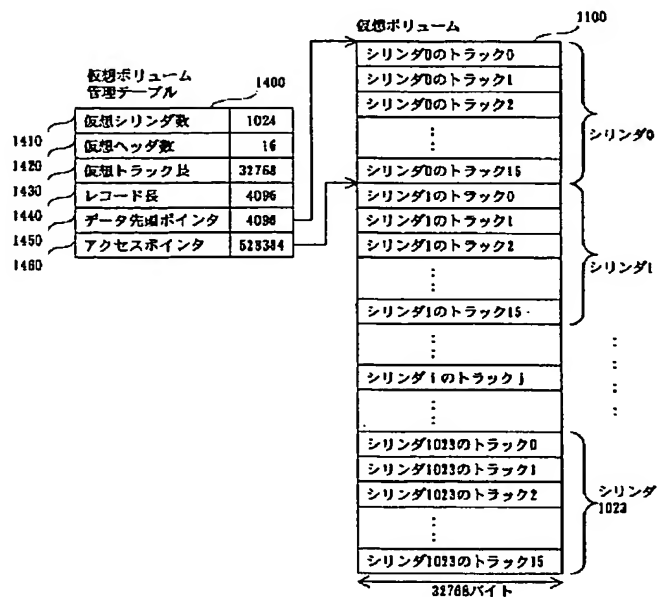
【図3】

図3



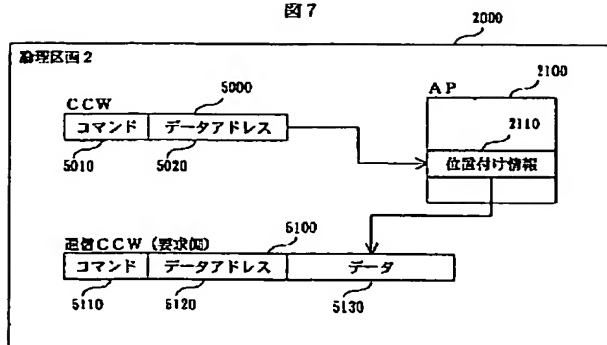
【図4】

図4



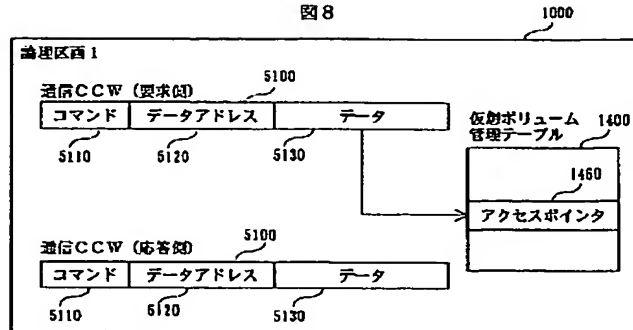
【図 7】

図 7



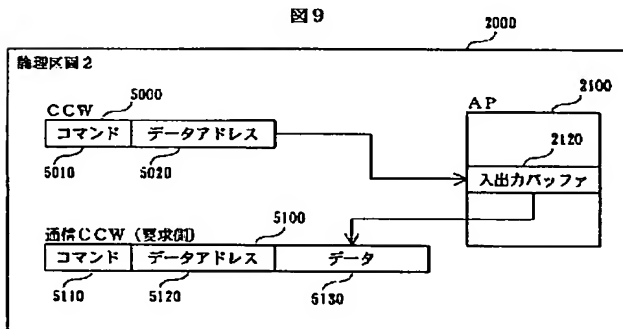
【図 8】

図 8



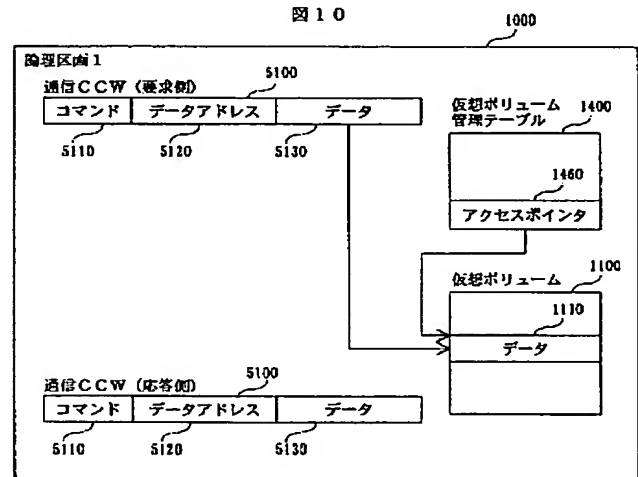
【図 9】

図 9



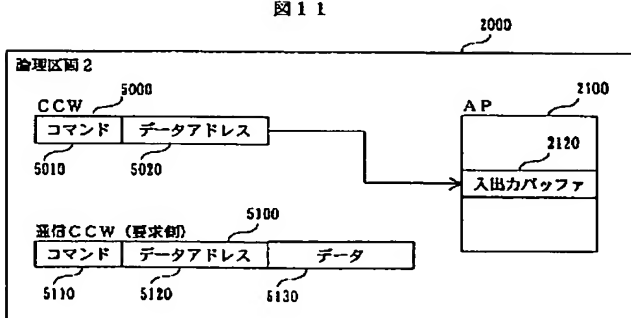
【図 10】

図 10



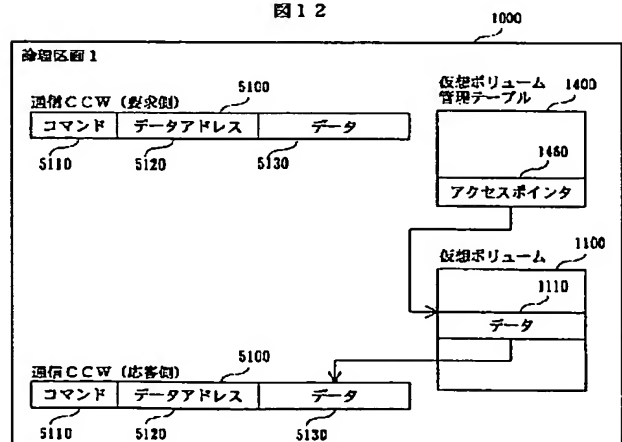
【図 11】

図 11

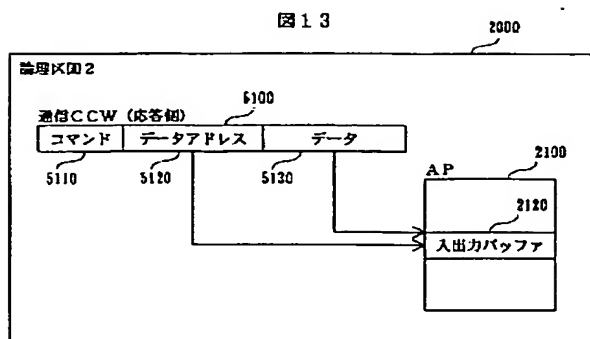


【図 12】

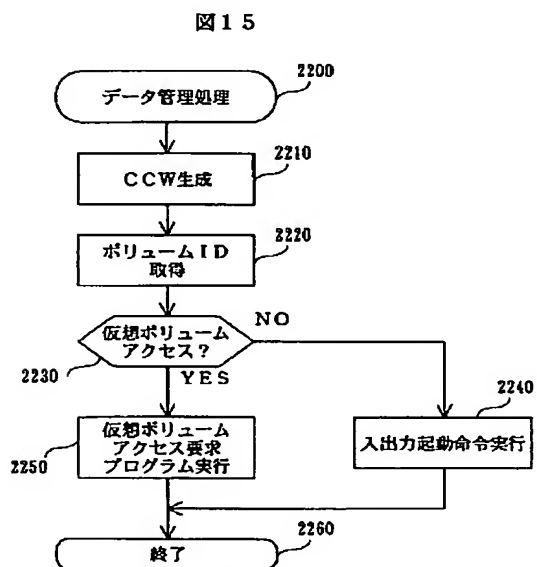
図 12



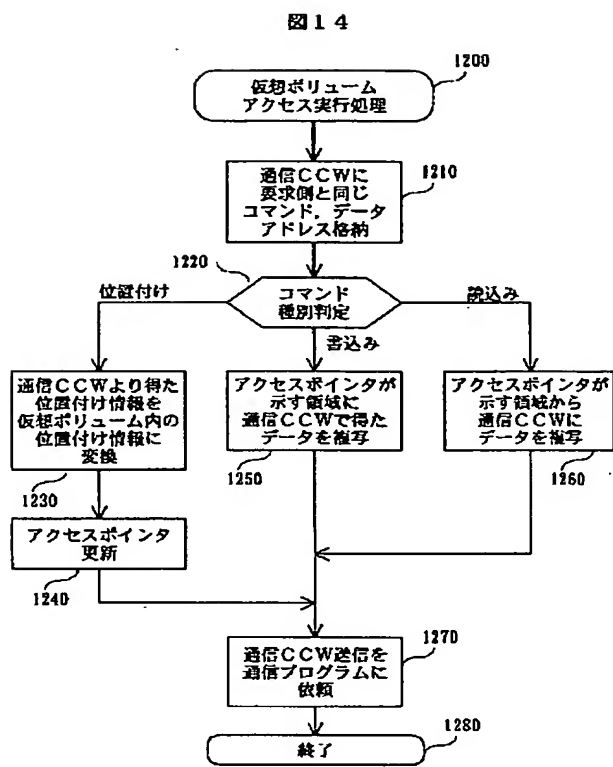
【図13】



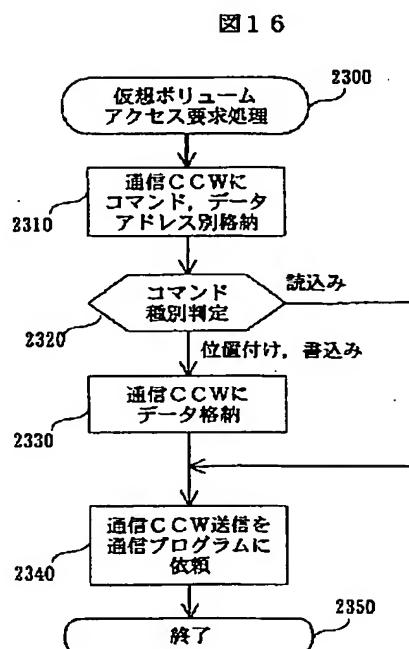
【図15】



【図14】

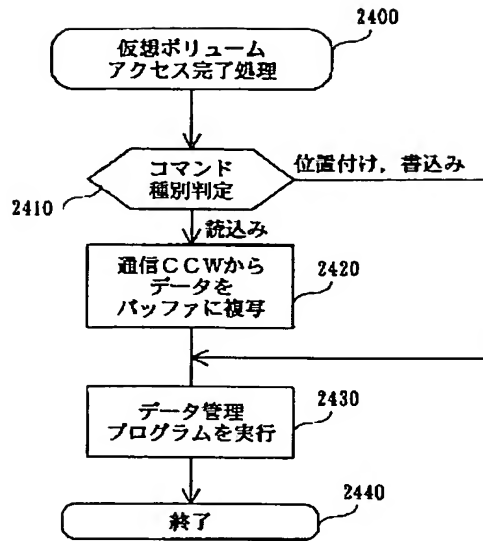


【図16】



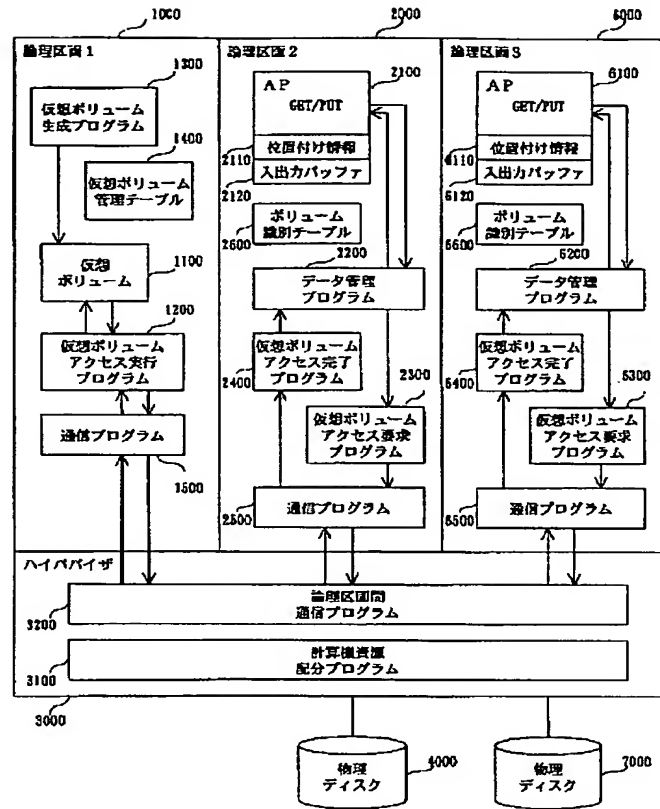
【図17】

図17



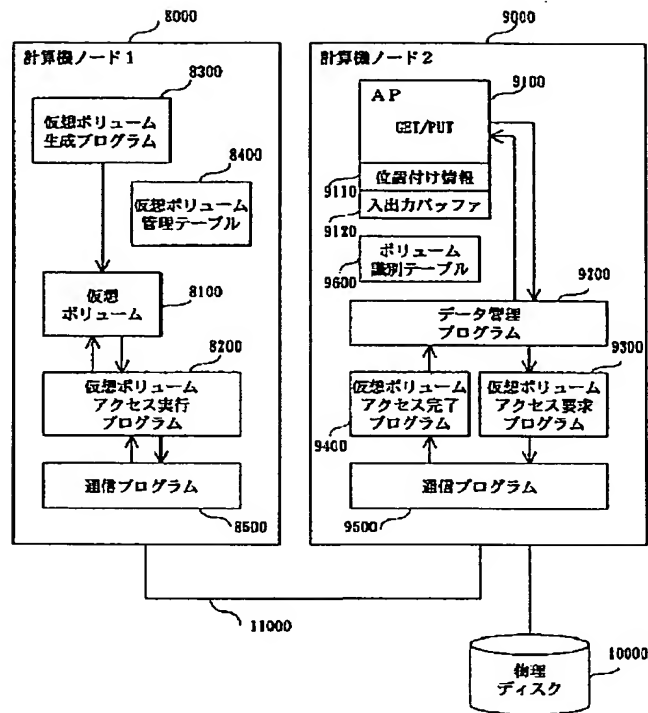
【図18】

図18



【図19】

図19



フロントページの続き

(72)発明者 大辻 彰
 神奈川県横浜市戸塚区戸塚町5030番地 株
 式会社日立製作所ソフトウェア事業部内
 (72)発明者 池ヶ谷 直子
 神奈川県川崎市麻生区王禅寺1099番地 株
 式会社日立製作所システム開発研究所内

(72)発明者 平岩 友理
 神奈川県川崎市麻生区王禅寺1099番地 株
 式会社日立製作所システム開発研究所内
 Fターム(参考) 5B065 BA01 CA02 CA13 CC02 CC08
 CH04 PA12
 5B098 AA03 GA01 GD03 GD15 HH04
 HH07